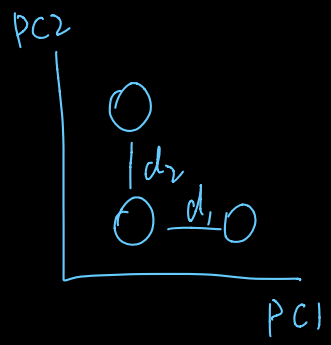
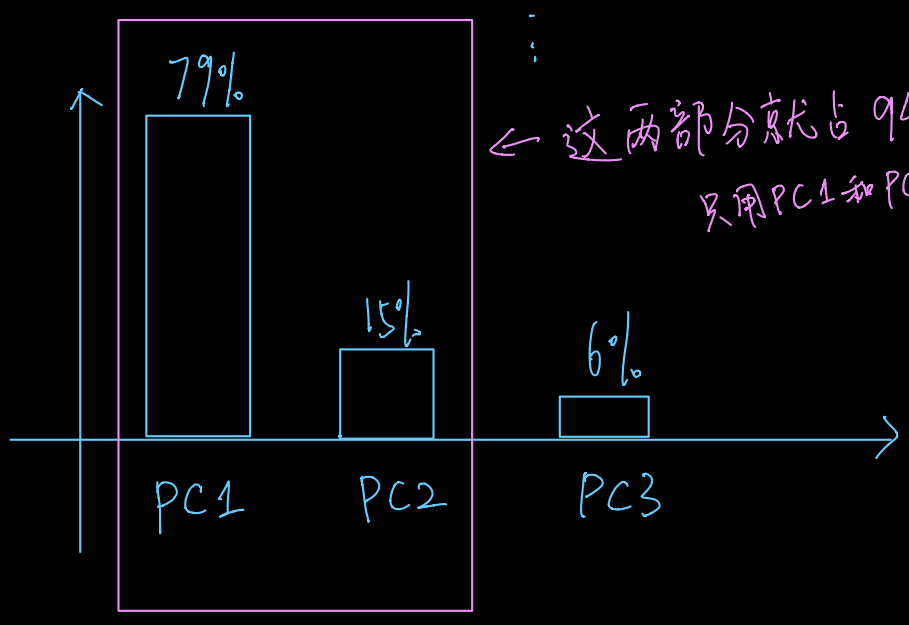
PCA:

- PC1 比 PC2 重要，所以

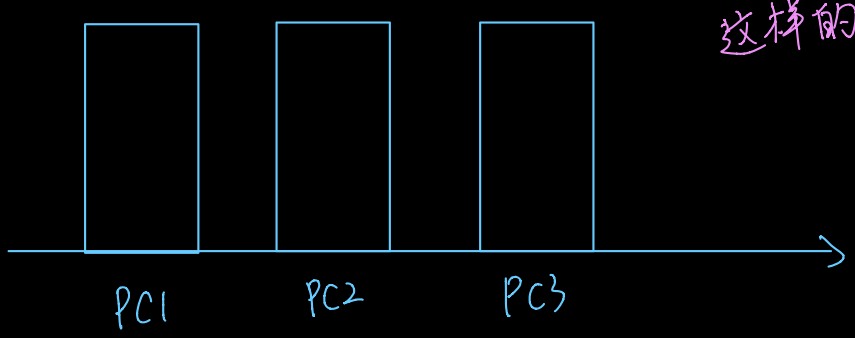


当 $d_1 = d_2$ 时，d_1 value more than d_2

- Variation for PC1 = $\frac{SS(\text{distances for PC1})}{n-1}$



← 这两部分就占 94%
只用 PC1 和 PC2 就可以了.



这样的就不能只用 PC1 和 PC2 了

PCA 代码:

```
pca ← procomp (traindata[, c(3:9)],
               center = TRUE,
               scale = TRUE)
```

goal: draw a graph that shows how the samples are related (or not related) to each other.

prcomp() return 3 thing:

- x — PCs for drawing a graph.

```
plot (pca$x[,1], pca$x[,2])
```

横轴 纵轴

- sdev — standard deviation, to calculate how much variation in the original data each PC accounts for.

```
library(ggplot2)
pca.data ← data.frame(Sample = rownames(pca$x), ← 1 col with sample ids
                      X = pca$x[,1],
                      Y = pca$x[,2]) } ← 2 cols for X, Y coordinates for each sample
```

data frame

```
ggplot(data = pca.data, aes(x=X, y=Y, label=Sample)) +
    geom_text() +  ← plot labels (text) rather than "dot" foreg.
    xlab(paste("PC1 - ", pca.var.per[1], "%", sep="")) +
    ylab(paste("PC2 - ", pca.var.per[2], "%", sep="")) +
    theme_bw() +  ← graph bg white
    ggtitle("My PCA Graph")  ← the title of the graph
```

- rotation

```
loading_scores ← pca$rotation[,1].
```

The prcomp() function calls the loading scores "rotation"

```
pcaScore ← abs(loading_scores)
```

最左边的有很大的负值.

最右边的有很大的正值.

abs() - sort based on the number's magnitude, rather than from high to low.

```
pcaScoreRank ← sort(pcaScore, decreasing = TRUE)

pca$rotation[pcaScoreRank, 1]   显示 scores (带+/-号)
```